ESTIMAÇÃO E PREDIÇÃO POR MODELO LINEAR MISTO COM ÊNFASE NA ORDENAÇÃO DE MÉDIAS DE TRATAMENTOS GENÉTICOS[1]

João Batista Duarte[2]*; Roland Vencovsky[2,3]

[2]*Escola de Agronomia - UFG, C.P. 131 - CEP: 74001-970 - Goiânia, GO.*
[3]*Depto. de Genética - USP/ESALQ, C.P. 83 - CEP: 13400-970 - Piracicaba, SP.*
**Autor correspondente <jbduarte@uol.com.br>.*

RESUMO: O presente artigo propôs-se a refletir teoricamente o processo de estimação/predição de médias de tratamentos, nos delineamentos em blocos, com ênfase nas suas aplicações em testes de genótipos, no melhoramento vegetal. Neste sentido, procurou-se comparar as análises baseadas no modelo linear fixo (análise intrablocos) e no modelo linear misto com genótipos aleatórios (análise recuperando informação intertratamentos), buscando identificar os fatores que podem determinar diferentes classificações genotípicas. A análise teórica permitiu constatar que a abordagem de modelo misto (com tratamentos aleatórios), comparativamente às análises tradicionais (médias marginais e análise intrablocos), em geral, leva a: *i)* maior homogeneidade das médias de tratamentos; e *ii)* seleção de diferentes tratamentos genéticos, quando a variância genotípica for baixa em relação à variância do erro e os ensaios forem não ortogonais e desbalanceados. Ademais, se os tratamentos forem oriundos de várias populações, a predição *BLUP* poderá determinar diferente classificação das médias de tratamentos, em relação à análise intrablocos, mesmo sob ortogonalidade e balanceamento.
Palavras-chave: recuperação de informação, delineamento em bloco, média BLUP, seleção genotípica, ordenamento

ESTIMATION AND PREDICTION USING LINEAR MIXED MODELS: THE RANKING OF MEANS OF GENETIC TREATMENTS

ABSTRACT: This study reviewed the theory of estimation/prediction of treatment means, in randomized block designs, emphasizing aspects of interest to plant breeders. Comparisons were made between analyses based on fixed (intrablock) and mixed (with random treatments effects - recovering intergenotypic information) linear models for identifying the determining factors that may affect the classification of genotypes. The mixed model approach, in comparison with the traditional analyses (marginal means and intrablock analysis), in general, leads to: *i)* more uniformly distributed treatment means; and *ii)* selection of different genetic treatments when the genetic variance is small relative to the environmental variance, as well as designs being non-orthogonal and unbalanced. In addition, if treatments of distinct reference populations are evaluated in the same experiment, *BLUP* prediction can lead to different ranking of means, in comparison with the intrablock analysis, even if designs are balanced and orthogonal.
Key words: information recovering, block design, BLUP mean, genotypic selection, shrinkage

INTRODUÇÃO

No melhoramento de plantas tem sido comum o uso de análise baseada em modelo fixo para a estimação de médias de tratamentos (ex: genótipos), mesmo quando estes foram obtidos por amostragem numa população. Isto é, em situações em que o modelo é tipicamente misto, pois inclui, além de efeitos fixos (ex: blocos), os efeitos aleatórios dos genótipos. Em boa parte dos casos, a modelagem mista é utilizada, com o rigor da suposição, apenas para a estimação de componentes de variância e para a construção dos testes F apropriados na análise da variância.

Entre as razões que levam os melhoristas práticos a não utilizarem predições baseadas em modelos mistos estão a falta de vivência com estes métodos e a sua pequena divulgação (Bueno Filho, 1997). Acrescenta-se que os efeitos prejudiciais da abordagem tradicional normalmente são tidos como mínimos, a ponto de não recompensar os esforços com a adoção da nova metodologia. A ordem de classificação dos genótipos, em geral, não se altera no caso de ensaios que seguem delineamentos ortogonais e balanceados. Assim, a estimação de médias admitindo-se modelo fixo, quando na verdade o modelo é misto, não modificaria o resultado final da seleção.

Por outro lado, a ocorrência de desbalanceamento não planejado, decorrente da perda de parcelas, é um fato normal nesse tipo de experimentação. Ademais, nas fases preliminares do processo seletivo, quando os genótipos são numerosos

[1]Parte da Tese de Doutorado do primeiro autor, apresentada à USP/ESALQ - Piracicaba, SP.

e ainda possuem natureza aleatória (Piepho, 1994), é comum o uso de delineamentos não ortogonais como *BIB* (*blocos incompletos balanceados*) e *PBIB* (*blocos incompletos parcialmente balanceados*). Também têm ganhado aplicação crescente os delineamentos aumentados (Federer, 1956), os quais, por construção, são desbalanceados e não ortogonais. Nestes casos, a possibilidade de classificações genotípicas diferenciadas entre as duas abordagens analíticas é uma realidade. Assim, optar-se pela conveniência da suposição de um fator como fixo ou aleatório pode estar longe de ser prática inofensiva (Bueno Filho, 1997).

Atualmente, a metodologia de modelos mistos tem-se tornado mais acessível aos usuários graças à sua implementação em sistemas estatístico-computacionais de ampla divulgação como o SAS® (Statistical Analysis System). Logo, a sua rigorosa aplicação é perfeitamente exeqüível sempre que o modelo subjacente aos dados for de tal natureza. Neste caso, covariâncias biologicamente conhecidas (ex: genótipos relacionados por origem e/ou parentesco) passam a ser levadas em conta não só nos testes estatísticos, mas também na estimação e predição de efeitos de implicação direta no ordenamento e na seleção dos genótipos. Os estimadores correspondentes, em geral, têm variância menor do que os de modelo fixo, resultando, assim, em estimativas de maior confiabilidade (Henderson, 1975; Verbeke & Molenberghs, 1997; Federer, 1998).

O propósito deste artigo é apontar, por meio de explicitações teóricas, os fatores que podem determinar diferenças na classificação das médias genotípicas (de tratamentos), quando estas forem obtidas por modelos fixo ou misto, de análise. O desenvolvimento centra-se na abordagem de modelos lineares mistos, por razões de generalidade e de divulgação. A ênfase principal está num modelo de delineamento em blocos, admitindo-se os efeitos de blocos como fixos e os de tratamentos como aleatórios. Ademais, procurou-se avaliar a extensão das constatações obtidas em algumas variações deste modelo.

UM MODELO DE DELINEAMENTO EM BLOCOS

Considere-se um delineamento experimental em blocos, com a tratamentos (genótipos) de efeitos g_i ($i=1,2,...,a$) e b blocos (completos ou incompletos) de efeitos b_j ($j=1,2,...,b$). Com o propósito de generalização, faz-se n_{ij} ser o número de vezes que o tratamento i aparece no bloco j ($n_{ij}=0,1,2,...$). Portanto: $\Sigma_i\Sigma_j n_{ij} = n$ (número de observações); $\Sigma_i n_{ij}=n_{.j}=k_j$ (tamanho ou número de parcelas do bloco j); e $\Sigma_j n_{ij}=n_{i.}=n_i$ (número de repetições do tratamento i); além de que $\Sigma_i n_i=\Sigma_j k_j=n$. Denota-se ainda por Y_{ijr} a observação num caráter ou variável aleatória Y (observável), relativa à *r-ésima* parcela ($r=1,2,...,n_i$) que recebeu o tratamento i, identificada também pelo bloco j. Um modelo linear que caracteriza esse conjunto de dados pode ser:

$$Y_{ijr} = m + b_j + g_i + e_{ijr}\ ;\quad \text{com: } e_{ijr} \sim N(0,\sigma_e^2);\ g_i \sim N(0,\sigma_g^2);\ E(Y_{ijr}) = m+b_j;\ \text{e}\ \ Var(Y_{ijk}) = \sigma_g^2 + \sigma_e^2.$$

Neste modelo, o efeito de bloco (b_j) é assumido como fixo e o de tratamento (g_i), como aleatório. A constante m é de natureza sempre fixa e e_{ijr} é uma variável aleatória não observável. Isso caracteriza o que se conhece na literatura por um *modelo misto*, pois incorpora uma mistura de tipos de efeitos, fixos e aleatórios (Searle, 1987). Dessa forma, os tratamentos testados representam uma amostra de uma população de genótipos, cujas respostas são distribuídas *normalmente*, em torno de uma média comum ($\mu_P = m + \bar{b}$) e com variância σ_g^2; ou seja, os tratamentos são realizações de variáveis aleatórias não observáveis, as quais correspondem aos efeitos g_i's (desvios genotípicos aleatórios em relação à média μ_P). O tratamento estatístico desse tipo de modelo, no campo do melhoramento genético vegetal, tem recebido ultimamente a denominação de *análise com recuperação da informação intervarietal* ou *intergenotípica* (Federer & Wolfinger, 1996; Wolfinger et al.,1997; Federer & Wolfinger, 1998; Federer, 1998).

Matricialmente, a expressão que generaliza essa e outras modelagens mistas alternativas pode ser escrita a partir do vetor $\mathbf{y}_{(nx1)}$ de observações, na forma do chamado *modelo linear misto geral*:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z\gamma} + \boldsymbol{\varepsilon}\ ;\quad \text{com:}\quad \boldsymbol{\varepsilon} \sim N(\phi,\mathbf{R});\ \boldsymbol{\gamma} \sim N(\phi,\mathbf{G});\ E(\mathbf{y})=\mathbf{X\beta};\ \text{e}\ Var(\mathbf{y})=\mathbf{V}_{(n)}=\mathbf{ZGZ}'+\mathbf{R}\ .$$

Neste caso tem-se: todos os efeitos fixos reunidos no vetor paramétrico $\boldsymbol{\beta}_{(px1)}$; os efeitos aleatórios no vetor paramétrico $\boldsymbol{\gamma}_{(qx1)}$, exceto os erros que compõem o vetor $\boldsymbol{\varepsilon}_{(nx1)}$; $\mathbf{X}_{(nxp)}$ e $\mathbf{Z}_{(nxq)}$ são as matrizes de incidências dos efeitos contidos em $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente; e $\mathbf{G}_{(q)}$ e $\mathbf{R}_{(n)}$ são as matrizes de variâncias-covariâncias dos vetores aleatórios $\boldsymbol{\gamma}$ e $\boldsymbol{\varepsilon}$, respectivamente, as quais compõem $\mathbf{V}_{(n)}$, a matriz de variâncias-covariâncias das observações. As covariâncias entre vetores diferentes são assumidas nulas (Henderson, 1984). Aqui, por simplificação, adotar-se-á: $\mathbf{G}=\mathbf{I}_{(a)}\sigma_g^2$ e $\mathbf{R}=\mathbf{I}_{(n)}\sigma_e^2$; onde $\mathbf{I}_{(.)}$ denota uma matriz identidade e $a=q$ (número de níveis do fator aleatório).

ESTIMAÇÃO E PREDIÇÃO NUM MODELO LINEAR MISTO

Sob as condições anteriormente definidas, o *método de quadrados mínimos ordinário* (*OLS*) não é mais um bom procedimento de estimação, pois assume a simples estrutura $\mathbf{V}=\mathbf{I}\sigma_e^2$, minimizando: $(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta})$. A recomendação recai, então, sobre o *método de quadrados mínimos generalizado* (*GLS*), o qual contempla qualquer estrutura não singular de $\mathbf{V}$, o que leva a minimizar a expressão mais genérica: $(\mathbf{y}-\mathbf{X\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X\beta})$. Todavia, para isso é necessário conhecer a matriz $\mathbf{V}$, através de $\mathbf{G}$ e $\mathbf{R}$, ou, alternativamente, inserir alguma estimativa de $\mathbf{V}$ no problema de minimização *GLS*. Nesta última situação, o que deve ser feito é encontrar razoáveis estimativas de $\mathbf{G}$ e de $\mathbf{R}$ por meio de algum método de estimação. Entre estes, destacam-se pelo volume de aplicações, os procedimentos *ANOVA* baseados no

método dos momentos (Fisher, 1918; Henderson, 1953) e os métodos de máxima verossimilhança, *ML* (Hartley & Rao, 1967) e *REML* (Patterson & Thompson, 1971).

Em várias situações, a preferência tem sido dada aos métodos baseados em verossimilhança, os quais exploram a suposição de que $\boldsymbol{\gamma}$ e $\boldsymbol{\varepsilon}$ têm distribuição normal (Littell et al., 1996; Verneque, 1994). Todavia, no caso de modelos mistos, não existe consenso sobre a melhor forma de estimar componentes de variância (Christensen et al., 1992). Optando-se por *ML* ou *REML* é necessário, então, construir uma função objetivo e maximizá-la em relação a todos os parâmetros desconhecidos. Segundo a abordagem do SAS Institute (1997), com alguns cálculos é possível reduzir o problema de maximização apenas aos parâmetros em **G** e **R**. Assim, os correspondentes logaritmos da função de verossimilhança (l_{ML}) e da função de verossimilhança restrita/residual (l_{REML}) são:

$$l_{ML}(\mathbf{G},\mathbf{R}) = -(1/2)\, log|\mathbf{V}| - (n/2)\, log(\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}) - (n/2)\,[1 + log(2\pi/n)] \; ; \text{ e}$$

$$l_{REML}(\mathbf{G},\mathbf{R}) = -(1/2) log|\mathbf{V}| - (1/2) log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - [(n-p)/2] log(\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}) - [(n-p)/2]\{1 + log[2\pi/(n-p)]\}.$$

onde: $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$; p, aqui, é o posto (*rank*) de **X**; e, dada uma matriz

A qualquer, $\mathbf{A}^{-}$ denota uma inversa generalizada de **A** (tal que $\mathbf{A}\mathbf{A}^{-}\mathbf{A}=\mathbf{A}$).

Do processamento numérico de uma destas expressões, através de algoritmos iterativos como Newton-Raphson (implementado no *PROC MIXED* do sistema SAS) ou *EM* (*Expectation Maximization*), pode-se obter as estimativas *ML* ou *REML* de interesse ($\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$). Nos outros métodos, a estimação fundamenta-se na construção de formas quadráticas do tipo $\mathbf{y}'\mathbf{P}_t\mathbf{y}$ ($t=1,2,...,s$; onde s é o número de parâmetros σ^2_t a serem estimados), as quais são equacionadas com suas esperanças matemáticas, $E(\mathbf{y}'\mathbf{P}_t\mathbf{y})$. As formas quadráticas equivalem às somas de quadrados obtidas na correspondente análise de variância e $E(\mathbf{y}'\mathbf{P}_t\mathbf{y})$ é uma função dos parâmetros σ^2_t. Descrições detalhadas destes métodos são disponíveis em Valério Filho (1991), Searle et al. (1992) e Lopes et al. (1993).

De posse dos valores paramétricos dos componentes de (co)variância (**G** e **R**) ou de suas estimativas ($\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$), passa-se, então, aos problemas de estimar o vetor de efeitos fixos $\boldsymbol{\beta}$ (ou uma função a ele associada) e de predizer o vetor de efeitos aleatórios $\boldsymbol{\gamma}$ (ou também alguma função de $\boldsymbol{\gamma}$). Ambos os problemas podem ser resolvidos, simultaneamente, através das chamadas *equações de modelo misto* (EMM), desenvolvidas por Henderson em 1948 (Littell et al., 1996; Henderson, 1984):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^0 \\ \tilde{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Não se conhecendo as matrizes **G** e **R**, simplesmente substituem-nas por $\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$. Manipulações de álgebra matricial levam, por conseguinte, às soluções do sistema:

$$\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \; ; \text{ e}$$
$$\tilde{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0) = \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$$

onde: $\mathbf{C}=\mathbf{G}\mathbf{Z}'$ é a matriz de covariâncias entre **y** e $\boldsymbol{\gamma}$ (covariância entre observações fenotípicas e valores genotípicos verdadeiros).

A obtenção das soluções do sistema por meio destas expressões não são, todavia, usuais, haja vista a dimensão *nxn* da matriz **V** a ser invertida. O uso de uma inversa generalizada da matriz de coeficientes em EMM representa uma opção de menor esforço computacional, uma vez que esta matriz tem dimensão *(p+q)x(p+q)*, inferior a *nxn* (McLean et al., 1991). Outras alternativas ainda são disponíveis (André, 1999).

É notório que, se $\mathbf{G}^{-1}$ tende para a matriz nula (ex: $\sigma^2_g \to \infty$, no caso particular $\mathbf{G}=\mathbf{I}\sigma^2_g$), as EMM tendem para as equações de *GLS* para estimar $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, quando os componentes de $\boldsymbol{\gamma}$ são considerados fixos (Robinson, 1991). Por outro lado, quando $\mathbf{G}^{-1}$ domina as EMM (ex: $\sigma^2_g \to 0$, sob $\mathbf{G}=\mathbf{I}\sigma^2_g$), $\tilde{\boldsymbol{\gamma}}$ tende para zero. Nos casos intermediários, $\mathbf{G}^{-1}$ opera reduzindo (*shrinking*, em inglês) a magnitude das estimativas de $\boldsymbol{\gamma}$ supostamente fixo, até zero (SAS Institute, 1997).

Se **G** e **R** forem conhecidas, $\boldsymbol{\beta}^0$ (ou, mais provavelmente, alguma função estimável $\mathbf{L}'\boldsymbol{\beta}^0$) é chamado *melhor estimador linear não viesado* (*BLUE - best linear unbiased estimator*) de $\boldsymbol{\beta}$ (ou de $\mathbf{L}'\boldsymbol{\beta}^0$), e $\tilde{\boldsymbol{\gamma}}$ é denominado *melhor preditor linear não viesado* (*BLUP - best linear unbiased predictor*) de $\boldsymbol{\gamma}$. O uso do termo *preditor* tem apenas o propósito de distinguir estimadores de efeitos aleatórios daqueles de efeitos fixos (Robinson, 1991). Porém, como já mencionado, **G** e **R** geralmente são desconhecidas, dispondo-se apenas de estimativas obtidas por algum método. Neste caso, os termos *BLUE* e *BLUP* não mais se aplicam, sendo apropriado substituí-los por *EBLUE* (*empirical best linear unbiased estimator*) e *EBLUP* (*empirical best linear unbiased preditor*), respectivamente (SAS Institute, 1997; Littell et al., 1996). O termo *empírico* é adicionado, portanto, para indicar esse tipo de aproximação.

A correspondente matriz de variâncias-covariâncias dos parâmetros, $\mathbf{C}_{\boldsymbol{\beta}^0,\tilde{\boldsymbol{\gamma}}}$ ou $\hat{\mathbf{C}}_{\boldsymbol{\beta}^0,\tilde{\boldsymbol{\gamma}}}$, é dada por:

$$\mathbf{C}_{\boldsymbol{\beta}^0,\tilde{\boldsymbol{\gamma}}} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-} \text{ ou } \hat{\mathbf{C}}_{\boldsymbol{\beta}^0,\tilde{\boldsymbol{\gamma}}} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix}^{-}$$

Dado que os resultados de partição destas matrizes são gerais, conhecendo-se ou não as matrizes paramétricas **G** e **R**, pode-se simplesmente escrever:

$$\mathbf{C}_{\boldsymbol{\beta}^0,\tilde{\boldsymbol{\gamma}}} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}'_{21} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \text{ ; com } \begin{cases} \mathbf{C}_{11} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \; ; \; \mathbf{C}_{21} = -\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\mathbf{C}_{11}; \text{ e} \\ \mathbf{C}_{22} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} - \mathbf{C}_{21}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \end{cases}$$

Note-se que $\mathbf{C_{11}}$ é a fórmula familiar da matriz de variâncias-covariâncias de $\boldsymbol{\beta}^{\mathbf{o}}$, solução de *quadrados mínimos generalizados*. Assim, entre outras propriedades, tem-se (Henderson, 1984; Searle et al., 1992):

$$Var\,(\mathbf{L'\beta^0}) = \mathbf{L'C_{11}L};$$

$$\text{Var}\,(\tilde{\boldsymbol{\gamma}}) = \mathbf{CV^{-1}C'\text{-}CV^{-1}XC_{11}X'V^{-1}C} = \mathbf{G} - \mathbf{C_{22}};\ \text{e}$$

$$Var\,(\mathbf{L'\beta^0} + \tilde{\boldsymbol{\gamma}}) = Var\,(\mathbf{L'\beta^0}) + Var\,(\tilde{\boldsymbol{\gamma}})$$

EXPLICITAÇÃO DO *BLUP* DE γ

Para entender as conseqüências da suposição de aleatoriedade dos efeitos de tratamentos sobre suas estimativas de médias é necessário primeiramente analisar o preditor $\tilde{\boldsymbol{\gamma}}$. É conveniente, portanto, derivar a expressão de componentes individuais do vetor $\tilde{\boldsymbol{\gamma}}$, isto é, de cada $\tilde{g}_i$, o *BLUP* de g_i (*i=1,2,...,a*). Sem perda de generalidade, omite-se, neste momento, o índice *j* (de blocos), mantendo-se apenas *i* e *r*, relativos ao tratamento e a sua repetição. Sob a estrutura de componentes de variância $\mathbf{G}=\mathbf{I}\sigma_g^2$ e $\mathbf{R}=\mathbf{I}\,\sigma_e^2$, tem-se: $\mathbf{V}=\bigoplus_{i=1}^{a}\mathbf{B}_i$ (onde $\mathbf{B}_i=\sigma_g^2\mathbf{J}_{(n_i)}+\sigma_e^2\mathbf{I}_{(n_i)}$; $\bigoplus_{i=1}^{a}$ indica a operação matricial soma direta, ou seja, a obtenção de uma matriz bloco diagonal com as *a* matrizes $\mathbf{B}_i$; e $\mathbf{J}_{(ni)}$ é uma matriz quadrada com todos os elementos unitários). Logo:

$$\tilde{\boldsymbol{\gamma}}=\begin{bmatrix}\tilde{g}_1\\ \tilde{g}_2\\ \vdots\\ \tilde{g}_a\end{bmatrix}=\mathbf{CV^{-1}}\begin{bmatrix}Y_{11}-\hat{Y}_{11}\\ Y_{12}-\hat{Y}_{12}\\ \vdots\\ Y_{1n_1}-\hat{Y}_{1n_1}\\ Y_{21}-\hat{Y}_{21}\\ Y_{22}-\hat{Y}_{22}\\ \vdots\\ Y_{an_a}-\hat{Y}_{an_a}\end{bmatrix};\ \text{com:}\ \begin{cases}\mathbf{C}=\bigoplus_{i=1}^{a}\mathbf{1}'_{(n_i)}\sigma_g^2\\ \mathbf{V^{-1}}=\bigoplus_{i=1}^{a}\mathbf{B}_i^{-1}\\ \mathbf{B}_i^{-1}=\dfrac{1}{\sigma_e^2}\left[\mathbf{I}_{(n_i)}-\lambda_i\mathbf{J}_{(n_i)}\right]\\ \lambda_i=\dfrac{\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2}\end{cases}$$

sendo: $\mathbf{1}'_{(ni)}$ um vetor linha (de ordem n_i) com todos os elementos unitários.

Portanto, a matriz $\mathbf{CV^{-1}}$, de ordem *axn*, assume o formato:

$$\mathbf{CV}^{-1}=\frac{\boldsymbol{\sigma}_g^2}{\boldsymbol{\sigma}_e^2}\Big[\overbrace{\begin{matrix}1-n_1\boldsymbol{\lambda}_1 & 1-n_1\boldsymbol{\lambda}_1 & \cdots & 1-n_1\boldsymbol{\lambda}_1\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0\end{matrix}}^{n_1\ colunas}\ \overbrace{\begin{matrix}0 & 0 & \cdots & 0\\ 1-n_2\boldsymbol{\lambda}_2 & 1-n_2\boldsymbol{\lambda}_2 & \cdots & 1-n_2\boldsymbol{\lambda}_2\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0\end{matrix}}^{n_2\ colunas}\ \cdots\ \overbrace{\begin{matrix}0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 1-n_a\boldsymbol{\lambda}_a & 1-n_a\boldsymbol{\lambda}_a & \cdots & 1-n_a\boldsymbol{\lambda}_a\end{matrix}}^{n_a\ colunas}\Big].$$

Por conseguinte, como ilustram Searle et al. (1992), o preditor do efeito genotípico de um tratamento *i*, ou seja, o *BLUP*(g_i), fica determinado por:

$$BLUP(g_i)=\tilde{g}_i=\frac{n_i\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2}(\overline{Y}_i-\mu^0)=n_i\lambda_i(\overline{Y}_i-\mu^0).$$

Isto resulta do produto da *i-ésima* linha de $\mathbf{CV^{-1}}$ pelo vetor $(\mathbf{y}-\mathbf{X\beta^0})=(\mathbf{y}-\hat{\mathbf{y}})$:

$$\tilde{g}_i=\frac{\sigma_g^2}{\sigma_e^2}\sum_{r=1}^{n_i}(1-n_i\lambda_i)(Y_{ir}-\hat{Y}_{ir})=\frac{\sigma_g^2}{\sigma_e^2}(1-n_i\lambda_i)\sum_{r=1}^{n_i}(Y_{ir}-\hat{Y}_{ir})=\frac{\sigma_g^2}{\sigma_e^2}(1-n_i\frac{\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2})\sum_{r=1}^{n_i}(Y_{ir}-\hat{Y}_{ir})\Rightarrow$$

$$\tilde{g}_i=\frac{\sigma_g^2}{\sigma_e^2}(\frac{\sigma_e^2+n_i\sigma_g^2-n_i\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2})\sum_{r=1}^{n_i}(Y_{ir}-\hat{Y}_{ir})=\frac{\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2}\sum_{r=1}^{n_i}(Y_{ir}-\hat{Y}_{ir})=\lambda_i\sum_{r=1}^{n_i}(Y_{ir}-\hat{Y}_{ir})=\lambda_i(Y_{i.}-\hat{Y}_{i.}).\quad (1)$$

Sabendo-se que: $Y_{i.}=n_i\overline{Y}_i$ e $Y_{i.}=n_i\overline{Y}_i$, tem-se, finalmente:

$$\tilde{g}_i=n_i\lambda_i(\overline{Y}_i-\overline{Y}_i)=\frac{n_i\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2}(\overline{Y}_i-\mu^{0_i})=n_i\lambda_i(\overline{Y}_i-\mu^{0_i})\qquad (2)$$

em que: $\overline{Y}_i=\mu^{0_i}$, representa a média dos valores ajustados (para os efeitos fixos) nas parcelas que receberam o tratamento *i*, ou seja, a média ambiental esperada naquelas parcelas.

Nos delineamentos em blocos, μ^{0_i} representa: (média geral) + (efeito médio dos blocos que receberam o tratamento *i*), o que pode ser demonstrado reintroduzindo-se o índice de blocos (*j*):

$$\mu^{0_i}=\frac{1}{n_i}\left(\sum_{j=1}^{b}n_{ij}m^0+\sum_{j=1}^{b}n_{ij}b_j^0\right)=\frac{1}{n_i}\left(n_im^0+\sum_{j=1}^{b}n_{ij}b_j^0\right)=m^0+\frac{1}{n_i}\sum_{j=1}^{b}n_{ij}b_j^0=m^0+\overline{b}^{0_i}$$

em que: m^0 e b_j^0 (*j=1,2,...,b*) são os elementos do vetor solução $\boldsymbol{\beta}^{\mathbf{0}}$, e $\overline{b}^{0_i}$ denota o efeito médio dos blocos que receberam o tratamento *i*.

Desse modo, o preditor do valor genotípico do tratamento *i*, o *BLUP*(g_i), pode ainda ser escrito como:

$$\tilde{g}_i=\frac{n_i\sigma_g^2}{\sigma_e^2+n_i\sigma_g^2}\left[\overline{Y}_i-\left(m^0+\frac{1}{n_i}\sum_{j=1}^{b}n_{ij}b_j^0\right)\right]=n_i\lambda_i[\overline{Y}_i-(m^0+\overline{b}^{0_i})]\qquad (3)$$

Este desenvolvimento mostra que o uso de uma constante μ^0, comum para todo *i*, como apresentam Searle et al. (1992, p. 271), não se aplica a todas as situações e delineamentos. Por isso, deu-se aqui preferência à notação μ^{0_i}. Mas, $\overline{Y}_i$ também estima tudo isso ($m+\overline{b}^{i}$) mais o efeito do tratamento (g_i). Assim, a diferença ($\overline{Y}_i-\mu^{0_i}$) contém, de fato, só a informação relativa ao efeito aleatório do tratamento *i*, ou seja, a média de seus efeitos genotípicos estimados por parcela. É também oportuno observar que o termo $n_i\lambda_i$ é equivalente à *herdabilidade de médias* de tratamentos ($h^2_{\overline{Y}_i}=\sigma_g^2/(\sigma_g^2+\sigma_e^2/n_i)$), conceito de ampla aplicação em genética.

Sob o ponto de vista experimental, é notória a importância de repetições e casualização para uma predição imparcial de g_i através de $\tilde{g}_i$. Embora isto também seja fundamental no caso de modelos fixos e seus respectivos *BLUE*'s, o raciocínio a seguir procurará ilustrar a relevância destes princípios no contexto da abordagem de modelos mistos.

Para contornar as dificuldades de um tratamento matemático através de esperanças condicionais, considere-se apenas as observações relacionadas a um dado genótipo *i*, cujo valor genotípico tenha sido predito por uma das expressões de $\tilde{g}_i$ (1, 2 ou 3). Para simplificação, considere-se ainda um delineamento binário, com $n_{ij}=0$ ou $n_{ij}=1$. Assim, pode-se afirmar que, sob as condições da predição, o valor esperado de cada unidade experimental é: $E(\hat{Y}_{ijr})=m+b_j$. Todavia, para esse

conjunto particular de dados (relativos ao genótipo *i*), g_i é uma constante, desconhecida, mas não uma variável aleatória. Logo, $E(g_i)=g_i$. Ademais, desconhecendo-se o número de repetições e as regras de alocação dos tratamentos às parcelas, não se pode, ainda, assumir: $E(e_{ijr})=0$; mas, sim: $E(e_{ijr})=e_{ijr}$. Neste caso, reintroduzindo-se o índice *j* (de blocos) na expressão (1) de $\tilde{g}_i$, tem-se:

$$E(\tilde{g}_i) = E[\lambda_i \sum_{r=1}^{n_i}(Y_{ijr} - Y_{ijr})] = \lambda_i \sum_{r=1}^{n_i} E(m + b_j + g_i + e_{ijr} - m - b_j)$$

$$E(\tilde{g}_i) = \lambda_i \sum_{r=1}^{n_i}(g_i + e_{ijr}) = n_i\lambda_i g_i + \lambda_i \sum_{r=1}^{n_i} e_{ijr}.$$

Sob n_i muito baixo e na ausência de casualização, $E(\tilde{g}_i)$ carrega um termo $\lambda_i \sum_{r=1}^{n_i} e_{ijr} \neq 0$. Isto prejudica a qualidade do preditor $\tilde{g}_i$ que terá um viés no sentido do efeito ambiental médio incidente nas parcelas que receberam o genótipo *i*. Todavia, sob casualização e n_i grande pode-se assumir, tranqüilamente, que $\sum_{r=1}^{n_i} e_{ijr} = 0$ e, nestas condições, $\tilde{g}_i$ resulta em predições não tendenciosas de g_i. Isto porque, garantidas as condições que tornam $E(e_{ijr})=0$, tem-se: $E(\tilde{g}_i) = n_i\lambda_i g_i$; e, sob n_i grande, $n_i\lambda_i \to 1$ (herdabilidade máxima), o que implica em $E(\tilde{g}_i) \to g_i$. Mas, sendo $0 \leq n_i\lambda_i \leq 1$, à medida que a herdabilidade diminui ($n_i\lambda_i \to 0$), o valor absoluto de $\tilde{g}_i$ reduz-se, proporcionalmente, no sentido do valor esperado populacional, $E(g_i)=0$. Isto revela o aumento da importância do relacionamento entre os genótipos, na predição do valor genético de cada um. Em síntese: $E(\tilde{g}_i) \to g_i$ quando $h^2_{\bar{Y}_i} \to 1$; senão, $E(\tilde{g}_i) \to E(g_i)$ à medida que $h^2_{\bar{Y}_i} \to 0$.

De fato, assumir $E(\tilde{g}_i)=g_i$ somente é desejável se o experimento der condições para tal, isto é, se for capaz de fornecer informações individuais em número suficientemente grande ($h^2_{\bar{Y}_i} \to 1$). Senão, à medida que a informação individual diminui ($h^2_{\bar{Y}_i} \to 0$), é preferível predizer a performance de cada genótipo atribuindo-se um peso crescente às informações de seus "parentes", ou seja, fica cada vez mais seguro admitir $E(\tilde{g}_i)=E(g_i)$. Assumir $E(\tilde{g}_i)=g_i$ (suposição do modelo fixo), nestes casos, implica num risco crescente de $\tilde{g}_i$ produzir estimativas pobres de g_i (parâmetro). Enfim, a abordagem de modelos mistos usufrui da flexibilidade de ponderar a informação individual, em detrimento daquela dos genótipos aparentados, conforme a confiabilidade associada à primeira. Já a metodologia de modelos fixos (*OLS*) não dispõe desta prudência.

O EFEITO "SHRINKAGE" NAS MÉDIAS *BLUP*

Inicialmente, convém introduzir o conceito de médias *BLUP*. Como o vetor $\boldsymbol{\gamma}$ contém apenas os desvios genotípicos associados aos tratamentos, para predizer, por exemplo, a resposta fenotípica de um genótipo *i* num bloco *j* é necessário construir uma função linear de parâmetros fixos e aleatórios: $\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^0 + \mathbf{Z}\tilde{\boldsymbol{\gamma}} \Rightarrow \tilde{\bar{Y}}_{ij} = (m^0 + b_j^0) + \tilde{g}_i$. Tal expressão representa: (valor médio do ambiente *j*) + (efeito do genótipo *i*). Contudo, se o pesquisador estiver interessado não apenas na informação dos efeitos genotípicos $\tilde{g}_i$ (suficientes para o ordenamento e a seleção de genótipos), nem em predições de parcelas individuais ($\hat{\bar{Y}}_{ij}$), mas na resposta média de cada genótipo, a expressão anterior não responderá ao seu questionamento.

A nova função deve levar em conta o efeito médio de blocos, assumido comum para todos os genótipos sob comparação. Isto corresponde às chamadas médias de tratamentos ajustadas para os efeitos fixos do modelo. A expressão é aqui denominada *BLUP* (μ_P+g_i) ou médias genotípicas *BLUP*, sendo dada por: $\tilde{\hat{\bar{Y}}}_i = (m^0 + \bar{b}^0) + \tilde{g}_i$. Esta, sim, determina o ajuste das médias de cada tratamento para um mesmo referencial, a constante ($m^0 + \bar{b}^0$) Computacionalmente, esse último termo é obtido construindo-se uma função linear dos efeitos fixos, $\mathbf{L'\beta}$, comum para todos os tratamentos. A matriz $\mathbf{L'}_{[ax(1+b)]}$ pode ter suas linhas todas iguais a: [*1* k_1/n k_2/n ... k_b/n], o que gera uma média ponderada dos efeitos de blocos pelos seus respectivos tamanhos, $\bar{b}^0$, à qual é adicionada a constante m^0. Acrescentando-se o preditor $\tilde{g}_i$ tem-se, então, a média de interesse.

Searle et al. (1992) tratam do problema de "estimar" ou "predizer" uma função linear do tipo: $\mathbf{w=L'\beta+\gamma}$. Os autores comentam que, para $\mathbf{L'\beta}$ estimável, $\tilde{\mathbf{w}} = \mathbf{L'\beta}^0 + \mathbf{CV}^{-1}(\mathbf{y} - \mathbf{X\beta}^0)$ tem propriedades de melhor preditor linear não viesado (erro médio quadrático mínimo, linearidade em relação a $\mathbf{y}$ e não tendenciosidade), sendo, por isso, chamado de *BLUP*($\mathbf{w}$): $\tilde{\mathbf{w}} = \mathbf{L'\beta}^0 + \tilde{\boldsymbol{\gamma}}$. No presente caso, tem-se: *BLUP*($\mathbf{w}$)= $BLUP(\mu_P + g_i) = \hat{\mu}_P + \tilde{g}_i = (m^0 + \bar{b}^0) + \tilde{g}_i$, o que corresponde à média *BLUP* do genótipo *i*. Littell et al. (1996) ilustram o tratamento desse tipo de problema através do sistema SAS. Os autores referem-se a tais combinações lineares como *funções predizíveis*, para diferenciá-las das *funções estimáveis* que combinam apenas efeitos fixos.

Definido o significado das médias *BLUP*, reconsidere-se agora a expressão de $\tilde{g}_i$. O termo $n_i\lambda_i$ (a herdabilidade de médias) representa um peso aplicado ao mais simples estimador de g_i, o desvio ($\bar{Y}_i - \mu^{0_i}$). Para avaliar a sua influência sobre este desvio, considere-se a relação $\phi_g=\sigma_g^2/\sigma_e^2$, a qual reflete a herdabilidade em nível de parcelas ($h^2_{Y_{ij}} = \frac{\sigma_g^2}{\sigma_e^2+\sigma_g^2} = \frac{\phi_g}{1+\phi_g}$). Assim, tem-se:

$$n_i\lambda_i = \frac{n_i\sigma_g^2}{\sigma_e^2 + n_i\sigma_g^2} = \frac{n_i\phi_g\sigma_e^2}{\sigma_e^2 + n_i\phi_g\sigma_e^2} = \frac{\sigma_e^2 n_i\phi_g}{\sigma_e^2(1+n_i\phi_g)} = \frac{n_i\phi_g}{1+n_i\phi_g}.$$

Numa situação de variabilidade genética muito superior à ambiental ($h^2_{Y_{ij}}$ de valor elevado) tem-se: $\phi_g \to \infty$ e $n_i\lambda_i \to 1$. Isto significa que a diferença ($\bar{Y}_i - \mu^{0_i}$) reflete, integralmente, o valor genotípico do tratamento *i* em relação à média μ_P da população, estimada por: $\hat{\mu}_P = m^0 + \bar{b}^0$. Nesta situação, a resposta média esperada do genótipo *i*, o *BLUP* ($\mu_p + g_i$), tende para: $\hat{\mu}_P + (\bar{Y}_i - \mu^{0_i})$. Em blocos completos balanceados, $\mu^{0_i}=\mu^0=\mu_P$; logo, sob

$n\lambda_i \rightarrow 1$, o *BLUP* ($\mu_p + g_i$) reduz-se a $\overline{Y}_i$. Neste caso, as respostas genotípicas obtidas pelo preditor *BLUP* ($\mu_p + g_i$) dispersam ao máximo entre si, igualmente às respectivas médias marginais simples não ajustadas ($\overline{Y}_i = \sum_{k=1}^{n_i} Y_{ij} / n_i$).

Por outro lado, quando essa relação de variâncias for muito baixa ($\phi_g \rightarrow 0$), o referido peso também diminui ($n_i\lambda_i \rightarrow 0$) e a diferença ($\overline{Y}_i - \mu^{0_i}$) pouco ou nada informará sobre o valor genotípico individual do tratamento *i*. Seja porque os tratamentos não diferem substancialmente entre si ($\sigma_g^2 \cong 0$), seja por erro experimental muito elevado ($\sigma_e^2 \rightarrow \infty$). Neste caso, a resposta média esperada de um genótipo *i*, o *BLUP* ($\mu_p + g_i$), tende para $\hat{\mu}_P$, pois, $\tilde{g}_i \rightarrow 0$; ou seja, todos os tratamentos terão respostas preditas idênticas ($\hat{\mu}_P$). Desse modo, variações fenotípicas observadas entre genótipos não são mais do que flutuações erráticas em torno da média populacional μ_P; pois suas propriedades genéticas individuais não são significativamente importantes ou, pelo menos, não puderam ser discriminadas pelo experimento. Nestas circunstâncias não haverá dispersão alguma entre as respostas genotípicas médias preditas. Com efeito, não seria aceitável qualquer variação entre estas numa situação de ausência de variabilidade genética. Logo, a abordagem de modelos mistos mostra-se coerente com a realidade e, por isso, é tida como conceitualmente mais completa (Resende et al., 1996; Bueno Filho, 1997).

Na maioria das situações práticas, entretanto, $n_i\lambda_i$ será um número entre *zero* e *um*, implicando num aproveitamento parcial da informação contida no desvio ($\overline{Y}_i - \mu^{0_i}$). Aproveitamento este proporcional à herdabilidade $h^2_{\overline{Y}_i}$, o que determina uma dispersão intermediária das médias *BLUP*. Portanto, conclui-se que uma redução na variância genética relativa (ϕ_g) implica num estreitamento da dispersão das respostas genotípicas médias preditas (Figura 1); podendo isto chegar ao limite teórico das médias se igualarem (quando $\phi_g = 0$). Trata-se do chamado efeito *shrinkage*, relatado na literatura de modelos lineares mistos, que nada mais é do que o "encolhimento" da distribuição das médias ajustadas de tratamentos em torno da média geral, quando se passa de uma análise assumindo-os como de efeitos fixos para outra em que tais efeitos são tidos como aleatórios. Quanto menor a herdabilidade, maior será o *shrinkage*. Este efeito é tido como uma propriedade desejável dos preditores (*EBLUP*, *BLUP*, *BLP*, *BP*), haja vista influenciar notadamente as médias $\overline{Y}_i$ extremas (Latour & Littell, 1996). Por razões desse tipo, tais preditores são também denominados *estimadores shrinkage* (Stroup & Mulitze, 1991).

Segundo Duchateau & Janssen (1997), em síntese, o *BLUP* representa uma contração da diferença ($\overline{Y}_i - \mu$) De forma que, se o componente da variância genotípica for bem maior do que o ambiental ($\sigma_g^2 >> \sigma_e^2$), o preditor $\tilde{g}_i$ será muito próximo de ($\overline{Y}_i - \mu$). Isto significa que a informação de outros genótipos relacionados não é muito útil para se fazer predições acerca do genótipo *i*.

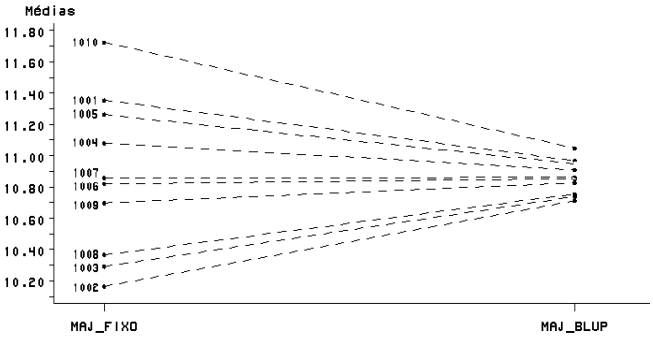

Figura 1 - Efeito *shrinkage* sobre médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP). Os números (1001 a 1010) identificam os genótipos, num ensaio simulado de blocos completos casualizados, sob: $\mu = 10$; $b_j \sim N(0, S_b^2 = 0{,}20)$; $g_i \sim N(0, \sigma_g^2 = 0{,}25)$; e $e_{ijr} \sim N(0, \sigma_e^2 = 2{,}00)$.

Mas, se $\sigma_g^2 << \sigma_e^2$, o preditor encolherá no sentido do valor esperado populacional (zero). Além disso, quanto maior o número de repetições (n_i), mais o valor ($\overline{Y}_i - \mu$) será considerado na predição individual.

Nota-se, portanto, que a abordagem *BLUP* é consistente com a intuição dos melhoristas de se suspeitar de um novo genótipo cujas respostas, em poucas repetições, têm média excepcionalmente alta ou baixa em relação aos demais (Hill Jr. & Rosenberger, 1985). Isto pois, a solução *BLUP* (à semelhança de outros preditores) leva em conta a informação de que os efeitos g_i têm menor variação do que as respostas dentro de cada genótipo *i* (Robinson, 1991).

ORDENAMENTO COMPARATIVO DAS MÉDIAS *BLUP*

Dado que no modelo em estudo μ_P é comum a todos os tratamentos (amostrados de uma mesma população), o ordenamento de suas médias preditas fica determinado apenas pelo de $\tilde{g}_i$: *ranking*[*BLUP*($\mu_P + g_i$)]=*ranking*[($\tilde{g}_i$)]. Por isso, para fins de seleção de genótipos, em geral, dispensa-se a obtenção das respostas médias preditas de cada tratamento. Em alguns estudos, todavia, o resultado das médias pode ser de interesse.

Da expressão de $\tilde{g}_i$ pode-se escrever: *BLUP* ($\mu_P + g_i$)= $\hat{\mu}_P + n_i\lambda_i(\overline{Y}_i - \mu^{0_i})$. Assim, dado que numa situação de blocos completos balanceados μ^{0_i} é comum para todo *i* e, inclusive, igual a $\hat{\mu}_P$, tem-se: *ranking* $(\overline{Y}_i - \mu^{0_i})$=*ranking* $(\overline{Y}_i)$. Ademais, sob balanceamento, o peso $n_i\lambda_i$ também é comum para todo *i* ($n\lambda$), implicando em: *ranking* ($\tilde{g}_i$) =*ranking*($\overline{Y}_i$). E, se a variância genética for bastante elevada ($\phi_g \rightarrow \infty$ e $n\lambda \rightarrow 1$), tem-se o resultado já obtido: *BLUP*($\mu_P + g_i$)= $\overline{Y}_i$. Por outro lado, se $n\lambda$ afasta-se de *um* (logicamente no sentido de *zero*) esta última igualdade não mais se verifica, embora a dos ordenamentos ainda permaneça, com a peculiaridade de *shrinkage* das médias *BLUP*($\mu_P + g_i$) em relação às médias $\overline{Y}_i$. Isto pois, à medida que $n\lambda$ tende para *zero*, a amplitude de variação

de $n\lambda\,\overline{Y}_i$ (termo determinante do ordenamento) reduz-se sensivelmente. Além disso, a constante $\hat{\mu}_P$ passa a ser multiplicada por (*1-nλ*), de valor também inferior à unidade. Logo: *ranking*[*BLUP*(μ_P+g_i)]=*ranking*($\tilde{g}_i$) =*ranking*($\overline{Y}_i$). Isto significa que, no caso de blocos completos balanceados, uma seleção baseada em médias marginais ($\overline{Y}_i$) levará à retenção e descarte dos mesmos genótipos que uma seleção baseada em *BLUP*(μ_P+g_i), ou em $\tilde{g}_i$ (Figura 1).

Resta, portanto, a questão de maior interesse prático relacionada ao processo de seleção de tratamentos: Existem ou não diferenças entre os ordenamentos de médias produzidas pela abordagem de modelo misto e por uma análise convencional intrablocos (modelo fixo)? A resposta é sim. Em blocos incompletos, balanceados ou não, μ^{0_i} não é mais comum para todo *i* e os ajustes para os efeitos de blocos podem fazer com que os ordenamentos dos genótipos por $\tilde{g}_i$ e $\overline{Y}_i$ não sejam os mesmos. Ressalta-se que a condição de desbalanceamento por si só (em qualquer delineamento) já é suficiente para não mais garantir a concordância perfeita desses dois tipos de seleção; sobretudo, se os tratamentos diferirem muito em números de repetições. A influência do nível de desbalanceamento sobre as médias preditas e seus postos pode ser avaliada diretamente na expressão de $\tilde{g}_i$ (3): Maior amplitude de desbalanceamento implica numa maior probabilidade de as duas classificações diferirem.

Independentemente do desenho experimental, do grau de desbalanceamento e dos números de repetições, quando $\phi_g \to \infty$ e $n_i\lambda_i=1$, tem-se: $\tilde{g}_i = 1$. $(\overline{Y}_i - \mu^{0_i}) = \tau^0_i$; onde τ^0_i é a solução do sistema de equações normais reduzidas da análise intrablocos, sob $\Sigma n_i\tau^0_i=0$. Em razão disso, as médias preditas da análise de modelo misto, *BLUP*(μ_P+g_i), serão iguais às médias ajustadas pela análise intrablocos ($\overline{Y}_{i(\text{ajust./fixo})}$). Mas, à medida que se afasta dessa condição limite ($\phi_g \to \infty$ e $n_i\lambda_i=1$), a igualdade entre $\tilde{g}_i$ e τ^0_i, bem como entre as médias correspondentes, não mais se verifica. E, embora a relação ϕ_g deva atingir a ordem dos milhares ($\phi_g \to \infty$) para uma igualdade quase absoluta das médias *BLUP*(μ_P+g_i) e $\overline{Y}_{i(\text{ajust./fixo})}$, seus respectivos ordenamentos podem permanecer idênticos mesmo sob relações bem menores. Na prática, valores de ϕ_g na casa das centenas, em geral, garantem coincidência absoluta das classificações. E, uma concordância razoável já é conseguida com valores de ϕ_g na casa das dezenas, o que resulta também em seleções muito similares (não obrigatoriamente idênticas) pelos dois procedimentos.

A relação ϕ_g não interfere somente no valor de λ_i, mas também no de $(\overline{Y}_i - \mu^{0_i})$, pois μ^{0_i} advém de $\mathbf{X\beta^0}$ e $\boldsymbol{\beta}^0 = (\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}}y$. Logo, na abordagem de modelos mistos, a solução para efeitos fixos também leva em conta a estrutura de variâncias e covariâncias das observações. Ademais, esta influência vai além da estimação pontual, interferindo também nos testes de hipóteses relacionados aos efeitos fixos, o que justifica sempre uma cuidadosa especificação da estrutura de erros (Littell et al., 1996; Duchateau & Janssen, 1997).

Diante disso, o costumeiro uso de análises fundamentadas na suposição de tratamentos fixos (ex: análise intrablocos) para fins de seleção de genótipos, quando, na realidade, eles forem aleatórios, desperta especial preocupação se a relação ϕ_g for baixa. E, sem dúvida, esse é o caso de boa parte dos ensaios de avaliação de genótipos em programas de seleção de espécies já bastante melhoradas, ou seja, com baixa variabilidade genética. É verdade que, nas fases preliminares do processo, a variância genética pode ser consideravelmente alta. Em contrapartida, nestas etapas, os erros experimentais, em geral, são elevados (grande número de tratamentos e pequeno número de repetições), implicando em baixos valores de ϕ_g. Nestes casos, o número de repetições (n_i) e o grau de desbalanceamento voltam a ter influência decisiva na ordenação dos genótipos pela abordagem aqui apresentada (modelo misto com tratamentos aleatórios). Isso porque, sob desbalanceamento, $n_i\lambda_i$ pondera diferentemente o valor do desvio $(\overline{Y}_i - \mu^{0_i})$ de cada genótipo, o que pode resultar em ordenações distintas dos tratamentos pelas médias *BLUP*(μ_P+g_i) e $\overline{Y}_{i(\text{ajust./fixo})}$ (ou por $\tilde{g}_i$ e τ^0_i, respectivamente).

DUAS OUTRAS VARIAÇÕES NO MODELO

Com base no desenvolvimento teórico apresentado procurar-se-á estender algumas constatações anteriores, sem demonstrações, a duas outras variações no modelo estudado. Na primeira, os efeitos de blocos serão admitidos como aleatórios ao lado dos de tratamentos (modelo aleatório). Na outra, os tratamentos (de efeitos aleatórios), são supostamente oriundos de diferentes populações, cada uma com propriedades específicas em termos de média e variância (ex: genótipos em estrutura de famílias).

Modelo com blocos aleatórios

A suposição de aleatoriedade para os efeitos de blocos e de tratamentos, num delineamento em blocos, corresponde à adoção de um modelo de análise com recuperação das informações interblocos e intertratamentos. Neste caso, apesar da modificação na estrutura da matriz **V**, o termo μ^{0_i} da expressão do *BLUP*(g_i) torna-se comum a todos os tratamentos ($\mu^{0_i} = \hat{\mu}_P = \hat{m}$), haja vista um único efeito fixo no modelo, a constante *m*. Logo, uma possível modificação na ordem de $\tilde{g}_i$, em relação à de $\overline{Y}_i$, dependeria somente dos efeitos diferenciados de análogos de $n_i\lambda_i$ (componentes de $\mathbf{CV^{-1}}$), pois: *ranking*($\overline{Y}_i$)=*ranking*($\overline{Y}_i - \hat{m}$). Sob pequena variação entre blocos, condição para o uso eficiente da informação interblocos (Malheiros, 1982; Kempton et al., 1994), tais efeitos aproximam-se de $n_i\lambda_i$, cuja influência no ordenamento das médias já foi anteriormente discutida.

Nos delineamentos em blocos incompletos (*BIB* e *PBIB*), a alta eficiência da análise com recuperação da informação interblocos requer uma relação $\sigma_b^2/\sigma_e^2<1/k$ (onde: σ^2_b é a variância de blocos e *k* é o tamanho dos blocos); ou ainda, um valor de $r<2$, onde: $r=1+k\phi_b$ e $\phi_b=\sigma_b^2/\sigma_e^2$ (Malheiros, 1982). Analogamente, resultados obtidos por Duarte (2000) sugerem que a eficiência do uso da informação intertratamentos exige relações ϕ_g inferiores ao inverso do número de repetições, o que implica, obrigatoriamente, em $\phi_g<1$. Logo, esta informação é especialmente importante quando a discriminação dos tratamentos torna-se dificultada pela baixa variabilidade genética (pequenos valores de ϕ_g). Esse fato aponta, mais uma vez, para os ensaios preliminares dos programas de melhoramento de espécies com uma longa história de seleção artificial, já bastante melhoradas e com pequena variância genotípica.

A recuperação da informação interblocos por si só, embora possa determinar trocas nas posições relativas das médias dos tratamentos (em relação às abordagens de médias marginais ou de análise intrablocos), não é responsável por *shrinkage* no conjunto das médias. Assim, uma possível maior concentração das médias de tratamentos obtidas a partir do modelo aleatório (médias *BLUP*), decorre do uso da informação intertratamentos.

Modelo com tratamentos de diferentes populações

Outro questionamento natural que surge ao discutir a ordenação das médias de tratamentos pelas abordagens de modelo fixo e de modelos mistos é: Como fica o ordenamento comparativo para um conjunto de tratamentos que são oriundos de diversas populações? Nos ensaios de melhoramento genético, os tratamentos podem representar diferentes linhagens ou progênies (genótipos) e as populações suas diferentes procedências, cruzamentos ou famílias. A análise de modelo misto aqui considerada, assume os efeitos de blocos e de populações como fixos e os efeitos de genótipos dentro de populações como aleatórios. Corresponde, portanto, a um modelo de delineamento em blocos com tratamentos hierarquizados em populações.

Dado que somente a abordagem de modelo misto utiliza a informação relativa às variabilidades genotípicas das populações, é possível surgir classificações bastante distintas pelos dois enfoques. Conforme já constatado, é de se esperar que as médias de progênies relacionadas a populações de baixa variabilidade genotípica apresentem valores próximos (*shrinkage*). Isto representa um mecanismo de agrupamento das estimativas de médias do qual a análise intrablocos (sob tratamentos fixos) não pode usufruir; haja vista não levar em conta a informação intergenotípica (Figura 2).

Este afunilamento das médias preditas quando se compara os dois enfoques, sobretudo para as populações de baixa variabilidade genotípica (ex: população P2, na Figura 2), pode determinar a troca de posicionamento relativo entre progênies de populações distintas, mesmo na presença de ortogonalidade e balanceamento. E, nas situações usuais de blocos incompletos, sujeitos a desbalanceamentos planejados ou não, esperam-se, inclusive, mudanças de classificações dentro da mesma população, o que pode, conseqüentemente, ter um forte impacto na seleção.

Finalmente, ainda se poderia perguntar: O uso dessa abordagem de modelos mistos não dificultaria a detecção dos chamados *segregantes transgressivos*, uma vez que há uma tendência dos $\tilde{g}_i$'s convergirem para o valor esperado populacional? A resposta é não. Primeiramente, porque, se as exigências da modelagem fixa forem satisfeitas, a de modelos mistos produz resultados equivalentes; mas, se não o forem, esta última reduz a chance de apontar genótipos comuns como transgressivos. Ademais, nesse tipo de abordagem não se pode ignorar a seleção intrapopulacional, concebida pela própria estrutura hierárquica do modelo. Os genótipos segregantes transgressivos caracterizam-se por valores de $\tilde{g}_i$ discrepantes em relação aos demais genótipos relacionados (da mesma população), podendo ser facilmente identificados. O melhorista deve, portanto, estar atento a este fato, praticando seleção entre e dentro das populações. Caso contrário, não estará explorando adequadamente os recursos da modelagem estatística menos restritiva.

CONSIDERAÇÕES FINAIS

É um equívoco admitir que na análise de um modelo com um fator aleatório, ao invés de fixo, apenas os componentes de variância (esperanças de quadrados médios) e os teste F podem se alterar. As constatações reforçam também a preocupação acerca dos problemas de especificação dos modelos de análise estatística na área do melhoramento genético vegetal.

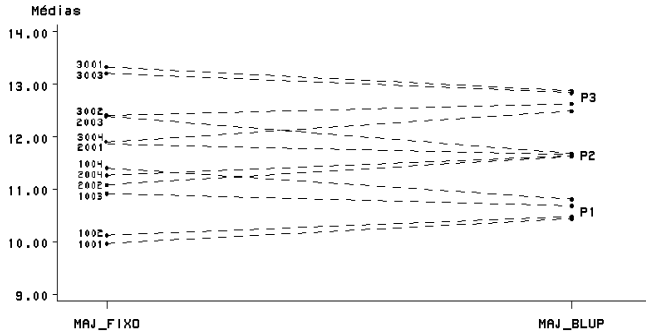

Figura 2 - Ordenamento de médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP), para tratamentos oriundos de três populações (P1, P2 e P3). Dados simulados para um ensaio em blocos completos casualizados, sob: $\mu=10$; $b_j \sim N(0, S_b^2=0,2)$; $e_{ijr} \sim N(0, \sigma_e^2=2,0)$; e $g_i \sim N(1, \sigma_g^2=0,15)$ se $i \in$ P1; $g_i \sim N(2, \sigma_g^2=0,05)$ se $i \in$ P2; e $g_i \sim N(3, \sigma_g^2=0,2)$ se $i \in$ P3.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos ao primeiro autor.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRÉ, C.M.G. Avaliação da melhor predição linear não tendenciosa (BLUP) associada ao uso de marcadores moleculares na análise dialélica. Lavras, 1999. 101p. Dissertação (Mestrado) – Universidade Federal de Lavras.

BUENO FILHO, J.S. de S. Modelos mistos na predição de valores genéticos aditivos em testes de progênies florestais. Piracicaba, 1997. 118p. Tese (Doutorado) –Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz".

CHRISTENSEN, R.; PEARSON, L.M.; JOHNSON, W. Case-deletion diagnostics for mixed models. **Technometrics**, v.34, p.38-45, 1992.

DUARTE, J.B. Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal. Piracicaba, 2000. 293p. Tese (Doutorado) - Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz".

DUCHATEAU, L.; JANSSEN, P. An example-based tour in linear mixed models. In: VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models in practice**: a SAS-oriented approach. New York: Springer, 1997. cap.2, p.10-61. (Lecture notes in Statistics, 126).

FEDERER, W.T. Augmented (or hoonuiaku) designs. **Hawaiian Planter's Records**, v.55, p.191-208, 1956.

FEDERER, W.T. Recovery of interblock, intergradient, and intervarietal information in incomplete block and lattice rectangle designed experiments. **Biometrics**, v.54, p.471-481, 1998.

FEDERER, W.T.; WOLFINGER, R.D. **SAS PROC GLM and PROC MIXED for recovering inter-effect information**. Ithaca: Cornell University, 1996. 8p. (Technical Report Biometrics Unit, BU-1330-M).

FEDERER, W.T.; WOLFINGER, R.D. SAS code for recovering intereffect information in experiments with incomplete block and lattice rectangle designs. **Agronomy Journal**, v.90, p.545-551, 1998.

FISHER, R. The correlation between relatives on the supposition of Mendelian inheritance. **Transactions of Royal Society of Edinburgh**, v.52, p.399–433, 1918.

HARTLEY, H.O.; RAO, C.R. Maximum-likelihood estimation for the mixed analysis of variance model. **Biometrika**, v.54, p.93-108, 1967.

HENDERSON, C.R. Estimation of variance and covariance components. **Biometrics**, v.9, p.226-252, 1953.

HENDERSON, C.R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, v.31, p.423-447, 1975.

HENDERSON C.R. Applications of linear models in animal breeding. Guelph: University of Guelph - Canada, 1984. 462p.

HILL JUNIOR, R.R; ROSENBERGER, J.L. Methods for combining data from germoplasm evaluation trials. **Crop Science**, v.25, p.467-470, 1985.

KEMPTON, R. A.; SERAPHIN, J. C.; SWORD, A. M. Statistical analysis of two-dimensional variation in variety yield trials. **Journal of Agricultural Science**, v.122, p.335-342, 1994.

LATOUR, D.; LITTELL, R. **Advanced general linear models with an emphasis on mixed models**: course notes. Cary: Statistical Analysis System Institute, 1996. 614p.

LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER, R.D. **SAS® system for mixed models**. Cary: Statistical Analysis System Institute, 1996. 633p.

LOPES, P.S; MARTINS, E.N.; SILVA, M.de A.; REGAZZI, A.J. **Estimação de componentes de variância**. Viçosa: UFV, 1993. 61p.

McLEAN, R.A.; SANDERS, W.L.; STROUP, W.W. A unified approach to mixed linear models. **The American Statistician**, v.45, p.54–65, 1991.

MALHEIROS, E.B. Efeitos da recuperação da informação interblocos na inferência estatística em ensaios em blocos incompletos equilibrados. Piracicaba, 1982. 110p. Tese (Doutorado) - Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz".

PATTERSON, H.D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v.58, p.545-554, 1971.

PIEPHO, H.P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. **Theoretical Applied of Genetics**, v.89, p.647-654, 1994.

RESENDE, M.D.V. de; PRATES, D.F.; JESUS, A. de. YAMADA, C.K. Melhor predição linear não viciada (BLUP) de valores genéticos no melhoramento de *Pinus*. **Boletim de Pesquisa Florestal**, n.32/33, p.3-22, 1996.

ROBINSON, G.K. That BLUP is a good thing: the estimation of random effects. **Statistical Science**, v.6, p.15-51, 1991.

SAS INSTITUE. **SAS/STAT® software**: changes and enhancements through release 6.12. Cary: Statistical Analysis System Institute, 1997. 1167 p.

SEARLE, S.R. **Linear models for unbalanced data**. New York: John Wiley & Sons, 1987. 536p.

SEARLE, S.R.; CASELLA, G.; McCULLOCH, C.E. Variance components. New York: John Wiley & Sons, 1992. 501p.

STROUP, W.W.; MULITZE, D.K. Nearest neighbor adjusted best linear unbiased prediction. **The American Statistician**, v.45, p.194–200, 1991.

VALÉRIO FILHO, W.V. Comparação de métodos para estimação de componentes de variância através de simulação de dados. Piracicaba, 1991. 160p. Tese (Doutorado) –Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz".

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models in practice**: a SAS-oriented approach. New York: Springer, 1997. 306p. (Lecture notes in Statistics, 126).

VERNEQUE, R.S. Procedimentos numéricos e estimação de componentes de variância em análise multivariada pelo método da máxima verossimilhança restrita – modelos mistos aplicados ao melhoramento animal. Piracicaba, 1994. 157p. Tese (Doutorado) – Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz".

WOLFINGER, R.D.; FEDERER, W.T.; CORDERO-BRANA, O. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. **Agronomy Journal**, v.89, p.856–859, 1997.

Recebido em 22.12.99